Autonomic Management of Next Generation Wireless Networks for Ubiquitous Services

(A White Paper submitted to NITRD Program, Aug 25, 2008)

1 INTRODUCTION

1.1 Background. Recent advances in the areas of Web services, mobile wireless communication and networking, smart devices, sensors and embedded systems promise interactions and services that have never been experienced before. As the advances in information and communications technology are riding Moore's Law, the ability to share services and resources has also witnessed exponential growth. In near future, such advances will lead to *ubiquitous service spaces* based on mobile and pervasive computing paradigm, that encompass data, computational resources, and services distributed over heterogeneous (wireline and wireless), pervasive network infrastructures. This will offer tremendous opportunities in a variety of novel and attractive application domains such as environmental monitoring and control, advanced automotive systems, critical infrastructure control, reliable healthcare systems, pervasive security, community based computing, social networking, cloud computing, and so on.

In these applications, typically a large-scale wireless networking system consisting of massive numbers of connected elements (e.g., devices, processors, sensors, and actuators) is designated to play a crucial role. The high complexity of such heterogeneous, dynamic systems raises new challenges in the science of design. These systems must be (i) *context-aware* and *self-adaptive* to uncertain internal and external environments to achieve reliable and robust performance, (ii) *self-managing* to prevent the escalation of maintenance cost, (iii) capable of *self-optimizing* the performance in different application scenarios, and (iv) capable of *self-organization* in order to enable rapid deployment and reconfiguration of the network infrastructure. This will enable the systems to respond to changes in the environment as well as the needs of the network operators, service providers and subscribers (application users).

1.2 Motivation. Current computing systems are rarely transparent and adaptive. They rely on human understanding of different computing platforms and technologies and often extensive manual programming to compose and aggregate services. This is clearly unrealistic in large-scale, highly unpredictable and dynamically evolving complex systems.

Uncertainty is the defining characteristic of mobile and pervasive computing networking systems and poses stiff challenges to the seamless functionality of ubiquitous services and applications. The uncertainty appears in various facets, for example at the physical layer (uncertainty in stochastic wireless channels and scarce spectrum), at the network layer (uncertainty in topology due to user mobility and wireless bandwidth availability) as well as at the application layer (uncertainty in traffic load and resource demands, application profiles and quality of service),. Today's mobile devices come with a number of embedded communication technologies such as 2G/3G/4G cellular networks, Bluetooth, Wi-Fi, WiMAX, cognitive radio, etc. As an example, in a disaster management scenario, rescue and public safety teams are likely to carry smart devices to communicate with each other. Some of the desirable requirements are the ability to transmit real-time voice, still pictures or low-quality videos, or even remotely actuate rescue equipment. Reliable communication and connectivity among the members of the rescue team play a significant role in the success or failure of the rescue operation and thus directly impact the lives that are in jeopardy. It is relatively easy to install a mesh or a peer-to-peer network to provide a wireless backbone. However, when a disruption occurs, the available wireless devices are unlikely to maintain a well-connected and reliable network. Instead, fractions or *clouds* of devices and network elements will be sporadically connected to each other, and possibly, to the surviving part of the infrastructure. In the extreme case, a single disconnected user may be a degenerate version of a cloud. Due to user mobility and spectrum scarcity, these clouds will be extremely dynamic. As a result, traditional networking approaches will fail to preserve dependable communication since many of them will require continuous end-to-end paths between communicating endpoints, and critical services based on multi-hop routing are unlikely to live up to the expectations. *Cognitive radio* permits dynamic spectrum assignment, and potentially alleviates part of this problem by opportunistically grabbing unused spectra and putting them to use at the time of crisis. Such systems must be spontaneously deployable, able to self-organize (so that nodes can seamlessly join or leave without the need of global control), reach consensus about maintaining the best-quality path in the face of conflicting requirements, and be able to self-heal following node or link or software failures. They must also be able to survive in spite of disruptive incidents and/or attacks by opportunistically exploiting and dynamically reconfiguring the available network components and resources. In other words, these systems must be *autonomic* whose holy grail is *self-management*.

Dynamic distributed systems go through periods of change and stability. An *adaptive system* reacts to changes in the *environment*, by encapsulating the needs of the users. Besides mobility and resource (e.g., bandwidth) requirement, the cognition and opportunistic self-allocation of unused spectra is an important part of the environment, which has interesting consequences: dynamically allocated parts of a spectrum may have to be released immediately after a designated user of that band appears in the scenario. This will trigger a disruption in communication, and a *self-healing* communication system has to restore functionality. Another significant feature is the need for *self-optimization*: a subset of nodes engaged in communication via a resource-constrained path must be able to spontaneously improve the communication quality in case resources like a specific band of the spectrum capable of providing better functionality becomes available, or a mobile node appears in the neighborhood promising to lower the latency. The scale of the application will have additional impacts on the functioning of such systems. A small-scale system deployed in a tornado-affected neighborhood will find it easier to abide by a set of policies established by a single administrator of the equipments. However, large-scale applications overseeing relief work in an earthquake-hit metropolitan area will have additional challenges: for example, the latency of some tasks may be unacceptable due to large number of routing hops. Also, the selfish behaviors of nodes under different administrative controls have the potential to adversely affect the overall performance in certain areas, as demonstrated by game theorists.

2 THE RESEARCH CHALLENGES

The diverse nature of the activities enabled by the next generation wireless networks and ubiquitous service requirements defines a grand challenge in computing whose investigations will require meaningful collaboration between research in various disciplines including mobile, and pervasive computing, autonomic computing, and adaptive service computing. Below, we highlight five fundamental challenges of next generation ubiquitous service spaces.

2.1 Challenge 1: Uncertainty Management

Uncertainty is the major driving force and perhaps the over-arching principle guiding an autonomic framework. Indeed, the uncertainty associated with the system produces unique challenges to achieving survivable communication services providing acceptable end-to-end quality of service (QoS). This calls for the design of new adaptive protocols to tame the uncertainty at various levels. Currently, uncertainties in wireless mobile networks are perceived mostly at the physical communication layer (due to rapidly varying wireless link qualities, or variable points of network attachment for mobile users). How to identify various uncertainties and their impact on the networking layers, as well as on the application goals? How to provide fault-tolerance and survivability when disconnections, disruptive incidents or attacks occur in an uncertain service support environment?

2.2 Challenge 2: Context-Awareness and Situation Modeling

Uncertainty management requires awareness of the contexts of the participating entities (such as devices or other system components), as well as the context of the entire environment or situation. A *context* is any relevant attribute of a device that provides information about its interaction with other devices and/or its surrounding environment at any instant of time. A sequence of device/entity contexts with the underlying interpretation (semantics) defines a *situation*. The understanding and analysis of the behavior of a system is paramount to capturing contexts unambiguously in the presence of uncertain (noise) and incomplete information. Past profiles help the system components anticipate future disruption and choose the best recovery action. How to generate *context awareness* and make the system *situation aware*?

2.3 Challenge 3: Autonomic Service Discovery

A crucial facet of self-management in mobile and ubiquitous computing environments is *service discovery*. Without this, nodes in infrastructure-less and dynamic environments such as mobile ad hoc networks (MANET) or vehicular ad hoc networks (VANET), will be confined to using only their own services and resources (or those that are pre-configured statically by system administrators). Even though many service discovery protocols and architectures have been proposed for such volatile environments, most are far from being considered autonomic. How to perform autonomic service discovery to cater to the needs of a volatile environment? How to utilize the promise of the newer tools like cognitive radios for efficient and expeditious autonomic discovery and recovery?

2.4 Challenge 4: Autonomic Service and Resource Management

Flexible and seamless configuration and delivery of services in large, autonomous, and complex evolving service spaces is another major goal. To achieve this objective, two strongly coupled aspects deserve attention: (i) composite services engineering, and (ii) resource provisioning. How to efficiently and dynamically manage the spectrum (bands) in cognitive radios? How to improve the QoS by preventing the over-provisioning of scarce resources (bandwidth) and using learning algorithms to profile and anticipate future resource usage, so that the resulting system performance is optimal or near optimal? How to facilitate efficient on-demand composite services configuration and protect users with guaranteed QoS despite variations in user activities, services spaces, and network infrastructures?

2.5 Challenge 5: Self-*Algorithms

The driving force behind successful self-management is the design of specific algorithms for different management tasks. How to design efficient algorithms for self-healing, self-optimization, self-stabilization, self-protection, fault-containment and graceful degradation to improve the availability and maximize the functionality of the applications?

While generic algorithmic tools for some of these are available, their applicability varies from one scenario to another. Accordingly, there is a pressing need to redesign them to suit the specific application needs. Two components are likely to improve the effectiveness of these algorithms. The first is the static data about uncertainly profiles collected from typical applications of the same class – this will help shape the ground rules or policies. The second is the learning mechanism using which the system becomes smarter with time: thus if the same fault repeats, or the same attack is launched several times, then ideally the system should be able to recover faster or protect itself more expeditiously with every new episode. How to add learning components to the various algorithms that are the cornerstones of autonomic computing?

3 CONCLUSION

Autonomic behavior of complex network systems consists of an assortment of *self-** (self-star) properties that guarantee dependability, availability and continued functionality. The proposed agenda will spawn further research not only in mobile/ pervasive and autonomic computing, but also in distributed algorithms, fault tolerance, security, game theory, and information theory. Eventual immunity to malicious actions is an issue that will continue to challenge the best minds as crooks invent new forms of abuse. The proposed framework must integrate different technologies (e.g., sensors, mobile devices, wireless mesh) prevailing in the (next generation) pervasive networking infrastructures. It will also be a crucial test bed for cognitive radio based applications that rely on the opportunistic harnessing of spectra. Ubiquitous (groupware) communication services will address real-time/interactive streaming (voice), interactive applications, mobile instant messaging, off-line streaming, chat, e-mail, data sharing, environment monitoring, and provide the assured quality of service that will strike a balance between security / privacy and usability.